Datapipelining for Heterogeneous Data Fusion

LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories

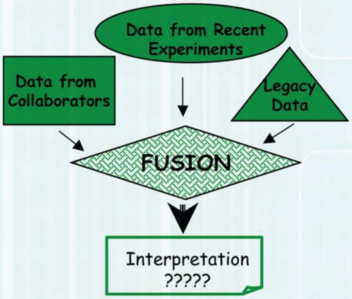
Genetha Gray, Pam Williams, Ken Sale, George Davidson, Brian Adams

other contributors: Mike Brown, Jean-Loup Faulon, Monica Martinez-Canales, Dave Gay, Keith Venderveen, Herbie Lee (UCSC), Matt Taddy (UCSC)

THE PROBLEM:

Draw conclusions given groups of related data sets

- Growing trend of collecting and interpreting large volumes of data
  - Genomics, proteomics, chemistry, medicine, …
- Different data types ⟶ different views of the same situation.
  - Multiple camera angles
  - Different experimental protocols
  - Human influence on data collection
- Seemingly dissimilar data may be complimentary.
- Data is often collected in a variety of formats.
- Data sets may vary in quality, quantity, adequacy, and relevance



OUR APPROACH:

Extend Ensemble Classification Techniques

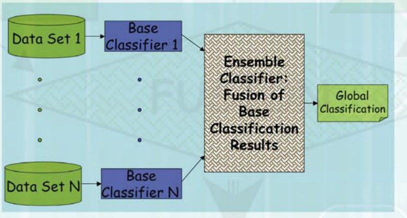
- Extract information from related but individual data sets and combine results.
- More accuracy than individual classifier.
- Typically apply different classifiers to the same data set. (Example: bagging, Breiman '96)

Advantages:

- Data can exist in separate databases.
- Does not require translation of data formats.
- Saves time and computational resources.

Areas of Research/ Challenges:

- Appropriate algorithms to oversee process.
- Demonstration of effectiveness for disparate data
- Specificity of applications



APPLICATION:

Protein Phosphorylation Site Prediction

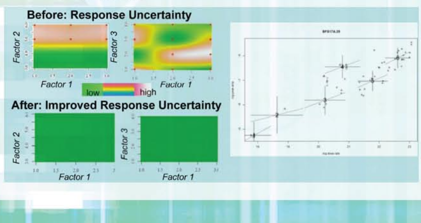
- Phosphorylation:
  - Most important regulatory cellular event
  - One focus of the MISL Grand Challenge
  - Sites are predicted using computational tools and verified experimentally
- Goal: Reduction the number of false positives
- 4 Base Classifiers:
  - NetPhos, KinasePhos, Scansite: use only the linear sequence of amino acids (open source)
  - Signature Descriptor: based on the molecular graph of the model (developed at SNL)
- Results: Intelligent combination of the classifiers reduced the overall number of false positves

Method	Serine (S)	Threonine (T)	Tyrosine (Y)	S/T/Y
NetPhos (n)	3837	1191	734	5762
KinasePhos (k)	1032	497	219	1748
Scansite (s)	1679	605	440	2724
Sig. Des. (d)	1989	790	390	3169
dkn ensemble	2100	581	408	3089
dks ensemble	1084	364	219	1667
kns ensemble	1626	517	358	2501
dns ensemble	1738	468	287	2493
dkns3 ensemble	919	237	178	1334
dkns2 ensemble	2355	728	458	3541
Sens ensemble	549	132	97	778
Spec ensemble	919	237	178	1334

APPLICATION:

Experimental Database Analysis

- **Activities:**
  - Demonstrate validity of simulation codes using a disparate set of experimental data
  - Use experimental data to improve computational model
  - Appropriately sample heterogeneous data to select representative points
- **Issues:**
  - Domain of interest is large but not all "testable"
  - Utilizing and supplementing exisiting data as appropriate
  - Data collected in different labs, has different formats, and contains noise
- **Solutions:**
  - Apply Bayesian statistical framework to improve the database quality for minimal cost
  - Development of new selection and sampling criteria to improve subsequent analysis



Sandia National Laboratories